

DOCUMENT RESUME

ED 093 972

TM 003 825

TITLE Proceedings of the Invitational Conference on Measurement in Education: Educational Measurement and Curriculum Change (Tallahassee, Florida, October 12-13, 1973).

PUB DATE 74

NOTE 26p.

EDRS PRICE MF-\$0.75 HC-\$1.85 PLUS POSTAGE

DESCRIPTORS *Conference Reports; *Curriculum Development; *Futures (of Society); *Measurement; *Testing

ABSTRACT

The speeches of John Dobbin and Irvin J. Lehman to the Invitational Conference on Measurement in Education are presented. In Lehman's paper, "Testing in the Year 2001", he states that testing in the future, or in any year, must provide data to assist students, parents, professional educators, tax payers, and legislators in the decision making process. Topics touched on are: technological advances in testing, including progress in optical scanners and equipment; sequential testing; item format; conceptual advances including more emphasis on testing in the affective and psychomotor domains; item matrix sampling; tab tests; tests to assist career guidance decisions; score reporting, criterion and norm-referenced tests. He also comments on such things as social and public concerns, invasion of privacy, and culture fairness. In Dobbin's speech, "Measurement and the Curriculum", he discusses what he considers some valid criticisms of the fact that curriculum is unrelated to reality. He predicts the end of standardized testing as we know it, comments on accountability in the future, criterion referenced testing, and postulates some possible avenues to pursue in the future regarding curriculum and evaluation. (RC)

ED 093972

003
822
TM

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

*International
Conference on
Educational
Research*

EDUCATIONAL RESEARCH AND CURRICULUM

London, Florida
October 1971-1972

Sponsors:

Florida A & M University
Florida State University
Leon County Board of Education
The Department, University of South Florida, Inc.

Patronized by:

Educational Area IV
Education Service
College of Education, F.S.U.

TM 003 822

Foreword

This conference was one of several which are sponsored annually by the Test Department Harcourt Brace Jovanovich and local educational organizations. The theme of the 1973 Gulf Coast Invitational Conference was "Educational Measurement and Curriculum Change." By most standards, the conference was a success. The sessions were well attended, there was considerable audience participation and a noticeable amount of positive feedback and correspondence has been recorded.

The conference sessions included small groups, panel discussions, a keynote speaker, a social hour, banquet and banquet speaker. Most of what took place in the small groups and panel discussions defies summary and reporting. However, both the keynote and banquet speakers have given permission for their presentations to be published. Their presentations are contained herein.

John Dobbin was the keynote speaker. He is a Project Director for Educational Testing Service and is well known to those in the measurement field. He helped establish and served as the acting director of the ETS Southern Regional Office in Atlanta during 1972-73.

Irv Lehman is a professor at Michigan State University where he works in the Office of Evaluation Services. He is also secretary-treasurer for the National Council on Measurement in Education.

No conference of this type takes place without the collective efforts of many people. Without their efforts there would have been no conference and, hence, no Proceedings. This writer wishes to express his thanks again to all who participated in making the conference a success.

H. W. Stoker
Conference Chairman

Testing in the Year 2001

Irvin J. Lehman
Office of Evaluation Services
Michigan State University

Up until about 1960, we in the field of measurement and evaluation were very complacent. We were complacent because the testing critics were yet to come. We were complacent because accountability, performance contracting, and testing moratoriums were still in the distance. But things ain't like they used to be back on the farm. We now have our various accountability programs, testing critics, moratoriums, legal hassles, and the like. Nevertheless, we still have much more that is right in testing than wrong. This doesn't mean that all is well and perfect--it isn't. But, if we can gain some insight as to what testing might be like in the future, I am confident that we will take the necessary steps to make teaching and testing even better than it is today.

Because it is nearly impossible to separate testing from teaching, some of my comments may appear to be related more to the instructional process, per se, than to the mechanics of testing or measuring the outcome of that instructional system. This, however, is intentional since many of the changes in testing that no doubt will transpire between now and 2001 will result from newer approaches to teaching. Also, I would like to point out that my remarks will be directed both to teacher-made as well as standardized tests although major emphasis will be placed on standardized approaches to testing.

I must confess at the outset that I really "dig" the topic since it gives me complete freedom to say whatever I want without being too concerned about the arguments that may come about when I am through. I think that you will all agree that my remarks are just my perceptions and that there can be no right or wrong....only that my perceptions are either very myopic or long-sighted.

In the year 2001, I really believe we will have more but better testing than now, the various moratoriums and testing critics yelling "foul" "racists" WASP--not-withstanding. Therefore, you and I need not become paranoid and fear for our future livelihood. Lehman (rather than Simon) says

"I see tests but not as they are now"

"I behold them but not as they are at present"

Testing in the year 2001--as a matter of fact, testing in any year--must provide data to assist students, parents, professional educators, tax-payers, and legislators in the decision-making process. In 2001, I envisage different kinds of decisions being made. For example, taxpayers will be even more demanding for accountability because of spiralling school taxes; students in ever-increasing numbers will be entering or at least considering highly skilled and technical but not professional careers; classroom teachers because of greater acceptance and use of IPI will be more concerned with formative as well as summative evaluation; and school board members and legislators will be vitally interested in performance contracting and assessment programs. And, because societal concerns and pressures in 2001 will be different from those facing us in 1973, I prognosticate BOTH different roles for, and different types of tests in 2001. I foresee an increase in the kinds

and amounts of testing in 2001 in comparison to what we have today. In fact, some of the worthy objectives of present education critics will best be realized by greater rather than lesser reliance on measurement and testing. Although it may appear in some of my prognostications that testing is doomed and rightly deserves to have a moratorium imposed on it, let me say loud and clear that education today and in the future is, and may suffer from a dearth rather than a deluge of testing.

What should be the role of measurement? Should it be as Schutz asks, "Servant, Soulmate, Stoolpigeon, Statesman, Scapegoat, All of the Above, or None of the Above?" As you will see in my ensuing discussion, any of these could be the correct answer. What we as professionals do in the next few years will determine the answer to the question.

I am told that any play worth its salt has one or more acts and scenes upon which the playwright develops his theme. Therefore, with humility, I present a NEW---3-Act play "The Role of Testing in the Year 2001." The outline for this production has been placed in your hands.

To present this 3-act New, thought-provoking, at times witty but nevertheless serious, play in the brief time allotted me and without the various stage props, dancing girls (for the men; fellas for the ladies) and musical accompaniment will be very difficult. But, being the masterful showman that I am and having such a good script to work with, I am certain at the 99% level of confidence that I will be successful.

The time at my disposal is not sufficient to more than sketch out a minimally functional set of procedural possibilities. Nevertheless, it may be of some use to use this occasion to focus our attention upon that part of the terrain that any fully orchestrated plan will occupy. We too frequently lose sight of the common ground, the fundamental across-the-board, parameters which should guide our common search for viable alternatives.

Even though I may be able only to sketch the highlights, I hope that I will leave you all so enthralled that you will prevail upon our hosts to ask me to return at some later date. As I always say, "Have speech; will travel; and the price is right."

In thinking about the various topics, any one of which could consume the time available, I attempted to organize my thoughts along the following points:

- A. What Have We Done?
- B. Where Are We Going?
- C. What Still Remains to be Done?
- D. Who Is Going To Do It?

And now, on with the play. Relax, enjoy yourself, and ask lots of questions when I'm through, even if they are irrelevant. Our hosts will be impressed and might ask me back again.

Technological Advances

The most dramatic changes in educational testing in the next 3 decades will come about because of increased technological progress in optical scanners and computers. By 2001, I see technology advanced to the stage that in one hour a scanner-computer will be able to score 40,000 papers each of which contains 20 subscores; perform an item analysis; transform raw scores to percentiles and other standard scores; my time limits (it may take 2-3 hours), I envisage every classroom teacher having such data on her desk overnight. Can you visualize how this rather immediate feedback will affect present-day teaching? Remedial instruction can begin within 24 hours after testing. Reteaching of

concepts, ideas, and knowledge, can begin before errors are firmly entrenched in the pupils' minds. With increased technological improvements we will be at the stage where data will be transmitted directly from the answer sheet to the computer. By 2001, we will truly be able to implement the IPI plan introduced in Winnetka in the 1920's but not popular then or now because of the tremendous burden it places upon the classroom teacher to prepare and score tests and to monitor the instructional system. I see, by 2001, computers producing and scoring tests; keeping a cumulative record of pupil progress; and containing decision rules to tell the student that his work needs to be reviewed, and, if so, which parts; or telling the student to proceed with this or that stage of the learning model. Such tests, of course, will be markedly different from our conventional standardized achievement tests which, by and large, are norm-referenced. Our newer tests will be content-specific and delimited to small segments of knowledge. Also, the information will be interpreted in a criterion-referenced system. Instead of answering whether Irv is better than Ruth (my wife, and of course he is) it will tell us whether Irv should continue studying and reviewing Unit III or whether he should move to Unit IV or V. Many psychometric problems are as yet unresolved but I foresee their solution by 2001. However, I do NOT see mastery testing as advanced by Bloom and others or item-generating rules and domain sampling such as espoused by Hively, Osburn, Popham and others being at the stage that they will be applicable to all subject matter fields. We are still lacking too much knowledge of the learning process to dully implement domain sampling.

Sequential Testing

Ideally, we should begin testing the examinee at a level where the item response will tell the examinee the point in the test where he can begin and yet safely assume that knowledge below that point exists. This, in essence, is the basal age concept in the Stanford-Binet. While we now have quasi-sequential testing in multi-level editions of standardized tests, practical problems of appropriate placement still exist for group tests to tell the examinee where to start and stop. But, with computer-assisted testing, decision rules can be stored in the computer's memory to carry out appropriate sequencing. I foresee standardized test publishers having computer facilities available so that any pupil can be tested at any time in any area without testing directors or counselors having to order tests beforehand. There will be a terminal in every school by 2001.

Item Format

The conventional answer sheet of the 1970's will 2001 be like the old Hollerith card is today -- a museum relic. Optical scanners in 2001 will be able to read and grade an essay; will be able to score responses made directly in the test booklet; will permit the examinee to cross out words or draw matching lines (thereby obviating the need for multiple-choice item formats that are so popular today), etc. I foresee different modes of presenting stimuli to examinees such as using recordings and pictures, as well as different response modes being used by the examinee. Although my latest remarks on new and novel item formats may appear to be contradictory to my forecast on greater computer use to generate and score test items, they need not be. In fact, I see that they are complementary rather than opposing thoughts.

CONCEPTUAL ADVANCES

Increased Emphasis on Testing in the Affective and Psychomotor Domain

In 1956, Bloom et al. developed the Taxonomy for Cognitive Objectives and today's tests now reflect this, judging from an analysis of the content validity reported in test manuals. In 1964, Krathwohl et al. developed the Taxonomy of Affective Objectives but little evidence is forthcoming that much use is being made of it. I hear also that work is progressing on the development of a Taxonomy on Psychomotor Objectives. I well recognize the unique problems associated with measurement in the affective and psychomotor domain but I see in my crystal ball, at least on the periphery, a concerted effort being made by both test publishers and classroom teachers to develop such tests. True, these tests might not resemble the familiar pencil-and-paper aptitude and achievement tests; they may not have written stimuli; they may not have the conventional answer-sheet response format. In 2001, we will have increased emphasis in our schools on the affective and psychomotor objectives, and both new tests and new ways of testing will have to be devised. I see us rising to the challenge!

Item Matrix Sampling

In large scale survey research and in assessment programs (which I will discuss later on) designed to provide answers regarding the "status quo" of the educational enterprise, it is ludicrous to administer the whole test or battery to all the students if we are only desirous of describing groups of students rather than individuals. Up until the last few years, our measurement techniques have focused on the individual as the object of measurement. Only since National Assessment and Project Talent have we begun to address ourselves to the development of measures which use a group of people as the object of measurement. Accountability as we know it today exemplifies the use of a group of people as the object of measurement whereas IPI and some types of performance contracting focus on the individual. I prognosticate that although group measurement will continue to grow in popularity -- sometimes for the wrong reasons -- it will complement rather than replace individual measurement.

Item by student matrix sampling permits the user to obtain large amounts of data in a relatively short time and at much less cost than total testing. It also permits us to obtain broader content sampling without subjecting any one pupil to an inordinate amount of testing. Because I see more accountability programs at both the state and local levels in the next three decades, I envisage greater use of this approach by test publishers. Ten years ago, we had no such thing as state or national assessment. In 1973, at least 14 states are involved in making comparisons of the scholastic achievement of their schools and pupils with national and regional data in an attempt to spot curriculum weaknesses. In the past two years, a large test publisher has established a State Assessment Center while some states and survey research contractors have formed consortia. Also, if the Quie bill presently being considered in Congress is enacted, it will force all states to undertake an annual assessment of scholastic achievement. This will involve testing, and no doubt, will involve item by student matrix sampling. This will involve the preparation of tests that are particularly relevant to a state's educational and instructional goals. Because there is still some uniqueness in the educational objectives among the states, I see test publishers becoming more involved in either producing directly or working with

state and local people to produce assessment-type tests and inventories. And, I see people like you becoming the interpreters and disseminators of test results.

Tab Test

Today's tests pay little attention to the measurement of "how the individual reasons." Although I foresee a movement in the next decade or two towards measuring the higher mental processes of synthesis and evaluation I also see, judging from what I hear today, that we will eventually return to a significant proportion of our achievement tests being factual recall. In other words, I see now, and the vision is growing stronger that parents are becoming agitated because their children are poor readers, poor communicators, and poor in the simple arithmetic operations. Therefore, although I see our test of 2001 emphasizing the total spectrum of learning, I still see a renewed interest in the measurement of "line of reasoning" and the decision-making process. I see new tests on the horizon. Present-day tests are not designed to do this. The tab test which is being used successfully in some of our medical schools can do this. However, we will need a different type of item format and we need research on how to score and evaluate the examinee's responses. But with the technological advances that I see in optical scanners and computers, I am confident that by 2001 the problems we now have in using the tab test to test large numbers of students will be resolved and that the tab test will occupy a role in the testing programs of senior high possibly junior high schools.

Tests to Assist in Career Guidance Decisions

I predict that career guidance will become much more important by 2001 because of the increasing number of students and parents who feel that a college degree no longer holds the aura and mystique that it once did. Also, because of changes in society, I see an increased need for skilled technicians to run our complex machines. It is true that there are many critics of testing in career guidance. However, as is true for any type of test, the criticisms levelled at testing are not solely due to poor tests. In the past few years we have witnessed advances in career guidance theory. Similarly, greater emphasis is being placed upon tests to aid the counselor and counselee in their decision-making process. As I said earlier, we have our testing critics. But it appears to me, at least, that blaming tests and only tests for invalid decisions is akin to blaming skin color for racial discrimination. The major purpose of testing in career guidance is or should be to provide information. Test scores, per se, do not tell us enough. The data--be they percentiles, raw scores, standard scores, etc.--must be transformed by the test author in collaboration with counselors. This transformation requires a sound theoretical orientation and empirical data. The test author supplies the former; the publisher, the latter. And yet, there are many test publishers, who in their eagerness to sell tests, place their products on the market without an adequate manual. I realize that some tests require longitudinal data to support their claim to validity, but if that is the case, the test hasn't proved its worth. Also, data about validity and reliability must accompany the test rather than have a note enclosed "Technical Data to Follow."

The counselor's role is to help change possibilities into probabilities. What better way (better in the sense of being more objective) is there to do this than by using empirically validated tests? Test scores

can and should be used to point out some of the possibilities and probabilities; test scores can provide clues as to how change can be brought about. The general public, even our minority and disadvantaged members, will realize that they have more to lose than gain by boycotting tests. But more on this a little later.

Disseminating Test Scores: New Reporting Services

Up until about five years ago, most efforts at obtaining and disseminating educational measurements have focused on the professional educator--be he teacher, counselor, school psychologist, or school administrator--and the decisions that he or she must make. Now, a few testing programs are addressing themselves to other decision makers--students and parent. Both of these are concerned with educational decisions. They may require different kinds of information. Parents want three types of information:

1. How well their child is doing in comparison to others in his grade.
2. How well their child is doing relative to the child's abilities.
3. How well their child is progressing towards major academic and personal goals established by the school.

Students, on the other hand, might be less concerned with knowing how well they are doing relative to their abilities. In any event, the conventional reporting services where percentiles, grade-equivalents, and stanines are reported presently will be supplemented and for these audiences, hopefully replaced by a detailed interpretative report. I predict that test publishers will computerize the process of presenting general information and that these programs will be different depending upon the audience to which the information is addressed. We have been remiss in providing educational decision-making information to school board members, taxpayers, and legislators. This, has created a hiatus which unfortunately is being filled -- oftentimes for the wrong reasons -- by state and local accountability programs.

Related to new reporting services, I still see a need for the personal approach to disseminating the meaning of test scores to parents and pupils. And because I predict that counselors of the future will occupy a more specialized role than they do today, we will need teachers who will be able to interpret and discuss intelligently the meaning of a test score be it obtained from a scholastic aptitude test, an achievement test, or a problem check list. Since I predict that such a course in 2001 will still be an elective in our teacher-training institutions, I must conclude that test publishers and people such as yourself will have to assume a greater role in providing in-service workshops. The future will demand that we have more and better interpreters and fewer advertising agents.

I see increased sophistication in the types of reporting services offered by test publishers. I see test publishers (some are already beginning to do this) providing not only test scores but producing individualized reports that are very specific and diagnostic in nature. I see test publishers, not just for the profit-motive but because of their genuine interest in furthering the aims of measurement, evaluation, and better teaching becoming very involved in the conduct of in-service training programs.

I predict more use of observational data-gathering devices such as rating scales, check lists, and anecdotal records. Some of these may be standardized. Many will be locally-made. You people here, I predict

will be the resource people for the preparation of, training in the use of, and interpretation of data from such instruments.

I predict that teachers of the future will demand an ever-increasing role in the decision making process and, therefore, I predict that they will rely very heavily on test scores and observational data.

Criterion and Norm-Referenced Tests

Criterion-referenced testing is the "IN" word in 1973. By 2001, I predict that it will still be part of the teachers' and possibly standardized test publishers' repertoire of materials, but there will no longer be opposing camps those for and those against criterion or norm-referenced testing. Criterion-referenced testing can play a vital role in education. It is ideally suited for formative evaluation and for those programs that stress individualized instruction. But I predict that the pendulum will turn so that emphasis will be on the use of norm-referenced tests. Why the lessened enthusiasm for C-R or mastery tests? Because people will come to realize that whether they like it or not, comparisons will always have to be made. Although we would like to think that all pupils are alike, we are deluding ourselves and venturing into Alice-In-Wonderland thinking when we think this way. When we need an electrician we don't think they are all alike! When we watch a football game we know that some players and teams are better than others. When we evaluate students we know that some are better than others. But where's our evidence? We'll have to get much of our evidence from tests and valid observational data. That's why I see greater reliance on tests, grades, and on norm-referenced tests. Again, let me say that there is and will be a definite role for criterion-referenced tests.

Finally, I see achievement tests - be they C-R or N-R - containing a subtest to measure writing ability. Our objective-type items used today are fine up to a point. But we must make a concerted effort to measure something more than a kind of inert, detached understanding.

Social and Public Concerns

Students, parents, and school personnel disagree on whether we are testing too much or too little; whether test scores really add appreciably to the making of valid decisions; whether tests are biased towards minority and disadvantaged students; whether testing is an invasion of privacy, and the like.

Invasion of Privacy

We are well aware that our best present-day scholastic aptitude tests account at most for about 55% of the variance in scholastic achievement measured by grades. Although the invasion of privacy issue has arisen primarily in relation to personality tests, I predict that future research pertaining to the prediction of scholastic achievement will have to consider biographical data, some forms of personality data, and some types of affective data--data which unfortunately will be deemed to be of a very personal nature. In addition, proper classification of students and prospective employees--and I predict that by 2001 our major thrust will be on classification rather than selection--will entail gathering personality-type and other kinds of data to help make better decisions. I predict that in 2001 people will still feel that their rights are being infringed upon. Therefore, I predict that testers and researchers will be more careful about the types of sensitive information

they gather and will be more sensitive to the concerns of students, parents, and prospective employees. Sensitive information will have to be gathered sparingly, judiciously, and used with discretion. But more important, the testers will have to empirically demonstrate that this information assists them in making more valid decisions.

A tangential problem of the invasion of privacy issue is related to the use of computers to store data gathered from the time a person is born until the day he dies. If my prediction on the increased use of tests to aid in formative evaluation holds true, and, if we will be gathering various kinds of affective data to make better decisions, this cornucopia of data can only be stored in a computer for easy access and rapid retrieval. Such storage raises a host of questions and problems. For example, what kinds of data should be stored? How long should data be stored? Who should be permitted access to the data? How will the rights of the examinee be protected? These are questions which by 2001 will result in data being available only to qualified school personnel and the examinee; data which will be up-dated frequently; and only those kinds of data that assist in the decision-making process will be stored. Unless we professionals begin to heed the cries of the populace and take some positive action, I predict that our courts will set down rules, which may protect the rights of the individual but may do more harm than good insofar as evaluation is concerned.

Culture-Fairness

I predict that when test scores are used to make valid decisions about an individual--whether these decisions relate to the amount of knowledge attained so as to sequence the pupil to the next appropriate learning task, or to help make valid predictions in terms of possibilities and probabilities of a person's future and not for screening purposes--the hullabaloo about culture-fair tests will diminish. Furthermore, I predict that unless we are able to gather the data necessary to conduct research, we will have a dilemma. In the past, in our attempts to bend-over backwards to avoid the taint of discrimination, we have not gathered separate data for different racial and religious groups. And this we MUST do. Otherwise, I predict that the problems we are presently facing in testing the disadvantaged and minority students will multiply to the extent that either (1) these members will refuse to be tested, or (2) we will become embroiled in a legal hassle which might put a lengthy moratorium on testing while the issue goes through the various courts, or (3) we will do a gross injustice to minority peoples, or (4) a combination of one or more of the above.

I predict that minority and disadvantaged spokesmen will realize that any type of moratorium will be more harmful than beneficial to their cause, even though the tests we use may not be perfect.

A Testing Moratorium

Regardless of what time period we look at, it is obvious that testing has been the scapegoat for the social and educational ills befalling our society. Students, teachers, professional educators and parents (minority and disadvantaged as well as the more affluent) have been screaming to declare a moratorium on testing. The moratorium was first directed to personality tests. Then it moved to scholastic aptitude tests. And now it is moving to achievement, particularly standardized achievement tests. I see a moratorium of some sorts in the immediate future but by 2001, it will be a matter of historical interest.

Students will realize that their goals are best served by data gathered objectively--the kind provided by well-constructed tests, rating scales, and check lists. I see the value of data gathered by means of observation--some of our most important educational objectives can be evaluated only with these kinds of data. But, from the standpoint of economy and efficiency, from the standpoint of objectivity, I predict greater rather than less reliance on tests.

Minority and disadvantaged spokesmen, I predict by 2001 will come to realize that their interests are best served by decisions made upon objectively-determined test scores. Tests, especially N-R tests must by their definition discriminate among individuals. I predict that more attention will be focused on the development of valid tests even though they discriminate (which they should). In any event, we must not ignore the screams of "foul" "racists" "culturally-biased." Rather, we must resolve to correct any existing deficiencies in our tests. I see by 2001 collaborative efforts on the part of all test constructors--be they teachers or test publishers--to involve minority and disadvantaged persons not only in the delineation of educational objectives and the preparation of test items, but also in decisions about the types of data to be reported and the manner in which the data are reported. I see a concerted effort being made to involve our critics in our search for better and different ways to obtain valid measurement tools. The NEW Stanford Achievement Tests have used minority and disadvantaged members in their planning and editing. NAEP has used them in the development and review of objectives and exercises.

Quality versus Quantity

The 7th MMY consists of two volumes and lists 1157 tests, many of them scathingly criticized. Now although the 1st MMY also contained some unflattering reviews, the point I wish to make is that we have witnessed a tremendous growth in the number of published tests and if the proportion of "rotten" ones remains constant, the number will increase. I predict that by 2001 we will have fewer but better tests than we do in 1973. Why? Because the public and the educational community will demand better tests. But NOW, you will ask will educators know the difference in quality between tests? They will have to be capable of making good judgments. I see professional educational decision-makers being more competent consumers because of increased emphasis by test publishers and people such as yourselves realizing that the misuse of tests will eventually, like Hamlet's ghost, come back to haunt us. That's why I earlier predicted that standardized test publishers and directors of testing will assume an active, on-going in-service role.

I have prognosticated about where testing will be in 2001. But how do we bring this about? Might I suggest that first we decide to work diligently and creatively at more meaningful dissemination. I see us repackaging test item responses and other input and output data according to the needs and sophistication of the user. Why report percentiles to parents or tax-payers? How meaningful are they? I also recommended that data be repackaged and ordered according to the type of decision to be made. For example, if the taxpayers in Dallas want to make local comparisons only, they don't need the kind of normative data usually supplied by the test publisher.

I suggest that we become more creative and flexible in the development and use of various techniques to generate, analyze, interpret, and

disseminate data. NAEP is a good example of how one might proceed. In music, they asked students to sing or hum a tune; to listen to a recording; to play a musical instrument. In citizenship, they had a discussion on various topics to see how decisions are made. I see our tests in 2001 resembling those presently in use in some respects, but markedly different in others.

EPILOGUE

Educators should be decision-makers! Regardless of the various reasons offered for testing in our schools -- "we've had tests since I came here in 1945," "it's on the school schedule," -- we must realize that testing is to provide information "to describe people in this or that trait...to predict behavior." Therefore we need relevant measures. I have uttered many predictions about testing in 2001. Regardless of whether you accept or reject my predictions, I think that you will all agree that testing in 2001 ain't gonna be like testing in 1973. The years between 1973 and 2001 may be turbulent from the perspective of educators and test publishers. But one thing is crystal-clear in my mind. We have done one hell of a good job and we can and will rise to the challenge of the future. To do a better job will require dedication, hard work, perseverance, a strong heart, and the ability to weather the winds of adversity which will come our way, sometimes in torando force. I am confident that we will rise to the occasion and weather the storm.

If I've succeeded in this brief and, I'm afraid I must admit, partial examination of what is really a knotty if not insolvable problem, I've accomplished at least a part of what I set out to do. If we get down to the fundamentals, take a clear look at the problems which confront us, now, we may be able to structure, to some degree, the appropriate mix of divergent foci by which we can bring to bear a large share of the energies and intelligence which reside within our profession. I would like to quote from Butler a statement on life which is applicable to our future in testing:

All our lives long, every day and every hour, we are engaged in the process of accomodating our changed and unchanged selves to changed and unchanged surroundings; living in fact, is nothing else than this process of accommodation; when we fail in it a little we are stupid, when we fail flagrantly we are mad, when we suspend it temporarily we sleep, when we give up the attempt altogether we die.

If we--you and I--are not prepared to face the challenge, we have no business in this vital educational role to which we have dedicated ourselves and to which we can contribute so much. At this point --the end of my Sermon-on-the-Mount a little prayer is in order.

O Lord

Give us the strength to accept with serenity the things that cannot be changed,

Give us the courage to change what can and should be changed,

Give us the wisdom to distinguish one from the other.

MEASUREMENT AND THE CURRICULUM

John E. Dobbin, Educational Testing Service

I have taken the topic of your conference as the title for my paper because, although this combination of words is as old as measurement, the phrase has taken on new meanings and a compelling new importance.

How so? What could have had a more thorough chewing -- no swallowing, just chewing -- than the tired old phrase: "Measurement and the Curriculum"? Let's start over and anticipate some of the new flavors that have been added.

I'll start with a generalization and then back up to flesh it out a bit. In general (I think), the curriculums of schools and colleges are one full generation behind the needs of the time -- and "measurement" is one full generation behind the curriculum. Making us T&M specialists two generations behind the world as it is. Working away at our computations of school means and average grade placements and standard deviations of never-mind-what, few of us are aware that the ideas central to our labors were fairly fresh when 23-Skidoo was "in."

There is no need for me to string together a lot of words to convince you that public school education lags behind the needs of our society. Enough books -- good books -- have been written about this lag to convince most of the reading public that "school" generally is a day-care operation with instructional content only occasionally and accidentally related to the demands that our environment places on young people.

The general theme of nearly all the critics (1,3,5), whether they are right and reasonable in the detail of their attacks or not, is that in the schools of today there are two main points at which curriculum is unrelated to reality:

- 1) The curriculum is not really related to the nature of the learner as a young human being, and
- 2) The curriculum is not really related to the nature of the environment with which the young human being must cope.

On the whole, I think that these criticisms are valid. And when a substantially irrelevant curriculum is applied by means of teaching methods that often are equally unrealistic, it is no wonder that the schools are targets for critical writers. So I am inclined to concede that the curriculum probably is a generation behind the needs of American society in these last decades of the 20th Century.

One intervening disclaimer before I plunge out of sight in my topic. Some of my older friends in the audience may recall that in an earlier incarnation I raved about the countryside prophesying that if the educators didn't break their shackles of tradition and teach kids what they need to learn in order to cope in this world, the business sector of our society would take the educating job away from the educators, and we'd all be out of jobs. I no longer believe this prophesy to be true. I publicly take it back. Having lately observed how capital B business goes about the job of educating when it has the chance, I note that capital B business has the same fatal flaw as capital E education -- that is, it knows all the answers before it starts -- and therefore is not likely to take over the education function. (So there go our chances of turning this terribly difficult job over to somebody else while we make our private fortunes in land development.)

Now to my task. First, I believe that "measurement" -- testing, evaluation, assessment -- as practiced in the schools almost unchanged since the 1920's will have disappeared by the end of the century. A few pockets of conservative practice will be holding out -- like the hill communities in West Virginia where Elizabethan English is still spoken, or the private schools, where kids are grouped in "forms" and refer to their teacher as "the Sir" -- but standardized testing as we know it now will have passed.

I'm not able to predict exactly what will replace standardized testing, but the direction of the coming change is already evident. Its impetus comes not from measurement specialists, nor even from curriculum people or administrators, but from school boards and legislators. It comes from people who ask in various ways for what we call "accountability" -- and who will be satisfied with no less than what we call "criterion-referenced" evidence of learning progress. Lets spend a couple of minutes on the ideas represented by those two terms.

"Accountability" is a relatively new term in the measurement world. We have split hairs for years over the distinctions among testing and evaluation and assessment, but accountability is different. It is related to the notion that surfaced during the last decade that educators should prove to somebody outside the brotherhood that they are doing what the society pays them to do. Actually, this idea was not created during the last decade; it existed long ago, but only recently has been reincarnated and given a new name. It is a nifty concept, not because it is any barn-burner in a technical sense, but because it may force us educators to say in

clear English what we are trying to accomplish with school children and then offer some convincing proof that we are accomplishing these things. It's as simple as that, in concept -- but its execution requires a whole new approach to our job by us measurement specialists. I'll get to that with an illustration in a minute.

"Criterion-referenced" testing is the other term I want to define, as I use it. To me, criterion-referenced testing is no more and no less than finding out if the student has changed in the way we wanted him to change. It means that instead of testing him on each small bit of input or practice exercise, we inquire into the effect this instruction and practice has had on him. For example, the golf pro, instructing a beginner, shows the student a whole raft of things to do and not to do in order to drive a ball straight down the fairway from a tee: feet so far apart, knees loose but not bent, left elbow straight on the backswing, keep your eyes on the ball -- and all the rest of it. The student practices each of these things time after time, with the instructor and by himself, until they come naturally and easily to him. But the criterion is driving the ball straight down the fairway from a tee. That's what pays off; that's what his instructor and his partners and his opponent's rate the learner on. He's terrific off the tee! His partners and opponents are applying a criterion-referenced test. He collects his wagers or loses his shirt according to his performance on that criterion-referenced test.

I have here a note to remind me to remind you to tell your friends in the school-marm business that, as the paired notions of accountability and criterion-referenced testing take hold in the education bureaucracy, they -- the teachers and administrators -- are going to be involved in assessment as they never have been involved before...involved every day, involved deeply, involved for the justification of their jobs, involved for the defense of their methods. Lest you conclude too quickly that here is just another old-timer with a Chicken Little syndrome, let me remind you of the Florida Education Finance Act of 1973. If you come from some other state, don't turn off your ears; this new law is likely to become a model for other states in the region. The Legislature, with the help of the Governor, has just passed into law a whole new system of state equalization of, and support for, education. This system of funding provides for every child in the poorest district as much educational money as is provided for children in the richest district. The people of Florida want this to be so and are willing to contribute enough additional funds in taxes to make it possible. In return, the Legislature and the people expect the professional educators to provide some understandable evidence that the financial investment of large additional sums of state money earns appreciable returns in terms of increased or improved educational outcomes. That much is in the 1973 law. It is my distinct impression, gained in conversations with some members of the education committees in this Legislature, that if the professional educators do not come up with that understandable evidence,

and soon, the legislators will take matters into their own hands, design accountability procedures that satisfy them, and write into law that school districts must meet a prescribed level of educational quality in output or suffer loss in state funds. This puts the local taxpayers on the side of educational accountability, and we all know what that means. (I must repeat, for emphasis, that it is my impression that the Legislature would take accountability into their own hands if they felt they had to in order to get quick action -- nobody actually told me this -- it is only a hunch.) But... no matter who finally calls the turns on this, criterion-referenced testing in a full system of accountability is something we will literally be forced to do, in Florida, by the elected representatives of the people. These things are no longer academic ideas to be rolled happily over the tongue, then put on the shelf when the game is over. They are the game.

There are some things in my own game, measurement, that I've never been able to understand. For example, by some process that defies reason, teachers and test-makers have conned themselves into believing that a statement like -- "Students in the fifth grade at the Sidney Lanier Elementary School earned an average (mean) score on the reading test in the Iowa Tests of Basic Skills that is three-tenths of a year, or 7 1/2 percentage points, above the national norm." -- constitutes understandable evidence about something. Even some parents have been brainwashed into believing that -- "Our school is nearly a whole year above the national norm" -- is a statement that has both meaning and usefulness. But our Legislature no longer believes these things; at least, some members don't. They say: "Hell...tell us what the kids can read, and how well." So there is our job.

And where do we start? You can start wherever you prefer -- or perhaps you have already started -- but I'm lazy enough to try first to use somebody else's work, if it is good, and go on from there. My preference as a starting point is the Henry Dyer monograph published by Phi Delta Kappa: How to Achieve Accountability in Public Schools. I won't repeat the details of the necessary assumptions in Dyer's hypothesis, but, as with any hypothesis with which we work, the assumptions on which it stands are vital to understanding of the outcomes of application. If you are interested in applying Dyer's technique, send PDK your fifty cents. But I will mention, for obvious reasons, that Dyer's method hangs on student change. How is the student different as a result of his school experience? For the purpose of finding out whether students change -- and if so how, in what directions, how much, and for what reasons -- Dyer suggest that INFORMATION of four kinds should be collected:

- 1) Input information. This information describes what the student is like "before" the school and the teacher and the curriculum go to work on him.

- 2) Output information. This information describes the student in comparable ways "after" the educational process has done its best with him -- thus giving us before-and-after information so that we may estimate change.
- 3) Process information. This information describes everything the school system does to produce change in the student -- the curriculum, teaching personnel and methods, extra-curricular activities -- all the human transactions that are supposed to help the student learn.
- 4) Information on surrounding conditions. This is information about the three kinds of circumstances that either help the student to learn or hinder his educational development: home conditions, school conditions, and community conditions.

As Dyer says, "To gather and organize and analyze the relationships among all four categories of information in ways which will show how well a school system is functioning in all of its segments for all of its students, and to show how the educational process in each segment might be changed to make things go better, is a task of enormous magnitude. Clearly it requires a high degree of cooperation among everyone who has anything to do with the schools: not only the students, teachers and administrative personnel -- but also parents, policy makers, and the public at large." (2; pp 29,30)

At this point I feel I must state an interjection. The Dyer monograph from which I quote so extensively is a technical book, describing the requirements of an hypothesis of accountability in its entirety. I don't expect -- and I don't think that Dyer expects -- that any school system, no matter how large or wealthy, ever will gather all the information it is possible to gather about all the students in all its schools and compute the thousands of multiple regression analyses needed to interpret relationships among all variables. Rather, it is my purpose here to describe the Dyer design as the theoretical frame within which I hope to suggest a few measurement procedures that are within reach of people who have to run schools, on the budgets and with the staffs available to them now.

Since our topic is "measurement and the curriculum," let's give curriculum a once-over-lightly in the light of the Dyer design. Almost all of the critics of education agree that a major function of the schools is that of creating a literate citizenry for the republic. By the time they get out of school, everyone believes, young people should know how to read and write and manipulate numeric symbols at least through the four fundamental operations. If any one of the Three R's is more important than the others, it's Reading. So, to make things easy for ourselves, let's concentrate on "measurement and curriculum in reading." We all know what the next questions are. How much reading skill is enough? How much of what kind of reading constitutes literacy?

Have you seen The Brief Test of Literacy, published in a descriptive monograph by the U.S. Public Health Service in 1968? (6) In order to establish a base point for the construction of a test of literacy for large samples of adults, the research team first used UNESCO definition of literacy, which expresses a generalization about "reading skill sufficient to permit active participation in the life of one's country." From that base, they went on to a definition of sufficient specificity to permit test development: "Literacy is that level of achievement which is attained by the average child in the United States at the beginning of the fourth grade." Literally, this means that the criterion is the reading skill attained or surpassed by only half of beginning fourth-graders! The beginning of the fourth grade! In this test, which was surprisingly difficult to construct because the fourth grade level of difficulty had to be achieved with item content that adults would accept, ^{its} difficulties were held within a narrow range and a YES-NO decision about the literacy of an individual (thus defined) could be made with a .91 reliability with seven reading passages and twenty-one questions, taking five minutes — clearly enough reliability for a survey instrument from which no individual interpretations will be made.

Okay...here's an assessment instrument related to a clearly-defined criterion. Notice, however, that the criterion itself is impure because it defines the goal as "the level of reading achievement which is attained by the average child in the United States at the beginning of the fourth grade." That is explicitly a norm — and we are in the odd position of having a criterion-based test but the criterion is norm-based! That point of departure is not bad, though, as these things go. I thought you'd like to know about this little instrument, if you don't already know about it.

But let's take another tack and set up some criteria for testing that have nothing at all to do with norms of any kind. Suppose, at the suggestion of some lawyers on a big city school board, we want to set up a curriculum to assure that 90% of the children in school will be able to "read the newspaper with understanding" by the time they reach the end of the sixth grade. Reasonable? Which newspaper? Wall Street Journal? New York Times? A tabloid? Suppose that we and our lawyer friends decide that a respected but widely-read newspaper in our area, like the Atlanta Constitution, should be selected as the source of criterion reading material. Now we hit the problem that even the newsstand dealer recognizes as a tough one in criterion selection: which part of the paper should our sixth-graders be able to read with understanding? The funnies? The editorials? Ann Landers? We decide that probably one lead story and one editorial — real ones, taken from back issues of the Constitution, not ersatz stories made up by English teachers — would make a good test. (Remember, this is not to be a test on which individual performances are interpreted and recorded, but a test to see whether the school and the teachers and the curriculum

are turning out human products that meet the requirements stated by the Board of Education.) So long as we have a narrow range of difficulty and only a YES-NO question to answer about each student tested, we can use a total of only five questions about this material and be almost 80% sure of our judgement about each student, attaining a reliability of .91 for our over-all conclusion. Actually, we could be almost as sure of our judgement about the student, and about our final conclusion, with only two questions: if the student gets them both right he passes. (These comments on the reliability of five-item and two-item tests, in carefully-described applications, are generalizations based on work of Frederic M. Lord that has been completed within the last month.) So...as long as we are willing to be as specific about the criteria of reading skill as I have been here in my example, we measurement blokes can track a student through his curriculum with assessment techniques we already have. If we were to use a series of two-item tests, we could locate in short order the kinds of things our sixth-grader can read and the things he can't. If the ancient superstitions of our craft won't let us go with two-item tests, let's just go along with five-item tests in which the decision on interpretation remains YES-NO, but the examinee is allowed to miss one answer out of five and still get a YES interpretation. So we can build satisfactory criterion-related tests of ability to do specific kinds of tasks in reading. In fact, if your school board insists, you can build a reading test of the kind I have just described -- make it fifteen minutes long and have it contain four articles from the local newspaper and ten questions, just to relieve the tensions among the old-timers who think that a good test has to be a long test -- administer it in those fifteen minutes, score the papers of every sixth grade class in the school in another hour, and by the end of the day let the superintendent report to his Board that 93% of the kids in their school system satisfy the reading skill requirement by the end of the sixth grade. He cannot tell them, from these tests, that the average sixth grader in the system is four tenths of a year better than the average in a national norming sample -- but he can tell them what proportion or number of kids in the grade have learned to do something that parents and citizens want them to learn to do. Simple, eh?

But now the plot thickens. There are two avenues of further inquiry that beckon. They beckon because you will look better if you think about them and explore them before those lawyers on the Board think of them.

- 1) Your tests will tell you how many kids, and which kids, are able to satisfy the reading criterion by the end of sixth grade. That is, they will tell you which kids can read stated kinds of journalistic material. What we don't know is which kids do read these parts of a newspaper with regularity. These reading skills don't do the individual or the community much good if the students do not acquire, along with their reading skills, the desire to put those skills to work. How about that? (This is the question I prefer not to pursue at this time.)

- 2) Second question: How do we know that all -- or most or some -- of the kids who read well enough to satisfy the criterion learned to do it in school? Should the teacher and the school and the curriculum get credit for pupil learning that was acquired somewhere else? Is it possible that the curriculum could have very little effect -- or even a negative effect -- even though 93% of the students knew how to read the newspaper by the end of grade six? (I don't prefer to pursue this question, either, but I promised to tackle one or the other, and this one is easier.)

You are now the superintendent. You have just told your Board with no little pride that 93% of the pupils finishing sixth grade in your system can read the specified kinds of newspaper content with understanding. Board members nod in agreement that this is good -- except one trouble-maker, who asks: "What produced this result -- the curriculum, the teachers, the teaching method, or the parents? Can the school system claim any credit for this?" On this question, it becomes apparent that measurement and curriculum and accountability and criterion-referenced testing and criterion-referenced teaching are woven together in a fabric of complex cause and effect that is demanding of tough, creative work among measurement specialists. What follows in the remainder of my paper is a description -- not of how I think you should do this tough, creative work -- but of how I think I would go about it if I had to.

First, I'd grab my copy of Dyer's Little Green Bible on Accountability. (This helps me to organize my thoughts about the problem before attacking it, which I find useful as well as reassuring.) It is reassuring in this case because on Pages 21 and 22 it says that this problem is a very difficult one to solve, so difficult that nobody has solved it yet. Obviously, we have a real one by the tail. It is also reassuring because on Pages 23 and 24 it says how difficult it is to find somebody to say what the criteria should be in a reading curriculum -- and with care that amounts to cunning we have avoided that problem simply by stating in our example that the criterion is a demonstrated ability to read with understanding one news story and one editorial from The Atlanta Constitution by the end of sixth grade. So a whole gerfluffle about behavioral objectives and what is criterion-related testing is behind us. (These things are easier to do when you are writing a book or giving a speech; when you come to something hard you just assume it.) All we have to do, to continue the example, is to find out how come more than nine out of ten of our kids pass this test at the end of grade six.

I've known a few superintendents who would be happy to leave well enough alone and allow the public to go on believing that the schools did it. But there is that one trouble-maker on the Board who keeps asking if this testing outcome is a result of the six hundred bucks per year that the taxpayers lay out for the paraphernalia of schools --

like books and buildings and teachers and curriculum consultants — or if it might be in part the result of some other factors operating in the community...so that kids in this town probably would know how to read this well by the end of grade six if the taxpayers laid out only three hundred dollars per year for each of the little darlins. This is where push becomes shove. And this is where the superintendent turns to his director of testing and evaluation and says, "Now it's your turn to earn your part of that \$600 per year!"

Returning to the Dyer reference, I am reminded that the kinds of information needed to answer a question like this are four: input information describing each student as he starts a phase of schooling; output information describing what he's like after that phase of schooling; process information describing what the school does to the child in the hope that he'll learn; and information about surrounding conditions that may have had an effect on the learning of many children.

I'd do the easiest things first. You might find out while doing the easier things that you don't need to do the harder ones. What are some of these easier things to do first? The very easiest are the ones that have already been done. Here are some of that kind:

- 1) Most schools and systems administer a test of "readiness" to youngsters in kindergarten or first grade. (Chances are better than 50-50 that it's an WBJ instrument.) Look up the class averages on the readiness tests of the class who have recently "passed" the criterion test to the extent of 93%. As a group, where was their readiness average in comparison with the group on whom the publisher normed that instrument? If their average when they started school was not appreciably higher than the average of the norming group, maybe the schools -- or something else that operated in the six-year period -- did put across some instruction that "took." It's only a hunch, but it's a start.
- 2) Next, get out the test records of this same grade group -- find their averages on every test of reading these kids took up through the sixth grade. What do the group's mean scores look like, when they are laid side by side? Are the mean scores in early years pretty nearly what you would have expected from their average readiness scores? Do the mean reading scores show some evidence of progression or growth from year to year? More progression than the norming group showed? If these changes do show up in the average scores on reading tests of the group under study, here is more evidence that what has happened to these kids in acquiring reading skill at least has happened during the period when they were in school, giving some face validity to the superintendent's claim that his schools produced good readers. So far, so good.

The two supporting actions I have suggested thus far are based on study of available data and can be done by a good clerk in an afternoon. Remember, we are trying to find out if that 93% literacy in the sixth grade has been the result of work by the school. The next thing I'd do is easy, too. I'd take the same test that showed the sixth grade to be 93% literate and administer it to the fifth grade, the fourth grade, and the third grade. This will provide some clues as to when this learning took place, though no clues as to where it happened. If 90% of the fourth-graders pass the same test that 93% of the sixth-graders passed, what have the fourth and fifth grade teachers been doing for two years that the kids should gain only 3% in two years (at six hundred dollars per year)?

Since the tests in literacy which I have hypothesized are only a few minutes in length and can be mimeographed at very low cost, I'll assume that you could talk a colleague in another school system -- another director of research or another director of evaluation or another coordinator of curriculum -- into giving the same criterion-related test you gave. It will require all kinds of grave promises and blood-oaths to get even a good friend to do this for you, though, because school administrators have a mortal fear of looking bad and don't take any chances that are not necessary. This kind of thing can be done, though, even if you have to get the help of some large city testing director who belongs to the secret order of large city testing directors who put on hooded robes and swap testing data from which all identifying marks have been removed. Once you have acquired performance data on your literacy test from a couple of other districts that you can describe but probably not identify, you'll have something new with which to support or refute your superintendent's claim. This kind of data smacks of comparison or "normative" information, of course, but it is the kind of information to which board members and taxpayers' associations attach some mystical significance, so add it to your modest heap of evidence.

Thus far, in discussing ways of proving that the schools have a vital role in teaching 93% of sixth-graders to read, I have talked only about supporting data from test tasks, administered to kids by teachers in school. And none of the information thus gathered -- I repeat: none -- contains any real proof of cause and effect. It offers clues, and nothing more.

So let's jump outside the test context in our next step and gather information by another means from another source. My choice of a proper jump here would be to ask parents where they think their children learned to read. I'd structure it a little more formally than the man-on-the-street quiz about what cleans toilet bowls best, but I would not make any big thing of it. (Make too big a thing of it, like building it into your doctoral dissertation, the first thing you know, you'll have some bright newspaper reporter move in with you, and you could have a very angry superintendent and no job.) So play it cool. Keep a low profile, as the government

burglars like to say. Put together a simple set of no more than half a dozen naive questions for parents, mimeograph them on recycled paper, and mail them to a sample of parents of sixth-grade kids. Be sure to enclose a stamped, self-addressed return envelope. In lots of places, the rate of return is much larger if the questionnaire goes out from, and is mailed back to, the child's own sixth grade teacher (How I miss that dear child!) than if it is identified as an inquiry from some high titled paper-pusher at the central office whose name the parents have never heard before. There are two aspects of this kind of inquiry to parents that require some expertise: the first being the design of the sample, which I assume all of you are quite able to do without any suggestions from me, thank you -- and the second is the phrasing of questions to be addressed to parents. On this latter point, I will offer advice, whether you want it or not.

School teachers and former school teachers are, by nature and definition, totally incapable of asking a question that a parent -- any parent -- can understand. So get somebody outside the school-main group to develop the questions for your inquiry. (When this admonition is rolled out, educators most often think of asking the editor of the newspaper or some famous novelist to draft their questionnaires for them. I have tried this and now think that the only two professional journalists I know who could do this are Norman Cousins and Art Buchwald.) So, with no budget and no access to teachers, I think I'd turn to the students -- the present sixth-graders. Gather a small committee of them -- interested ones, bright but not necessarily schoolish ones, reasonably representative of the cultural backgrounds found in the schools, not more than seven or eight of them in any one group, but every student in any group at least knowing all the others in his group to say "Hi!" to -- and I ask them: "What do you think we ought to ask your parents in order to find out how they think you learned to read so well?" (Since I'm an ex-teacher I cannot even ask this question properly. It will have some other words in it, but you get my meaning.) After they get over the shock of being asked for their advice, the sixth-graders will feed back -- not the questionnaire items themselves -- but clues as to what the content and mood of the questions ought to be. The questionnaire you develop from such sources will resemble the census-takers list in no way. I would advise you to boil it down to the extent that it will go on one side of an 8 1/2 by 11 sheet, double spaced, with a very short explanatory paragraph, no more than five or six questions that can be answered YES or NO, and plenty of space for writing in additional comments. Period. No more. And what you'll get back will be a set of fairly easy-to-interpret statements describing the school's impact on their children in the complicated business of learning to read -- from the special point of view, biased perhaps but usually interested, of the parents of the learners. Summarized in the simplest possible way, this information can do a great deal to support or negate the assumptions which the superintendent has made about his schools.

Now, a final step in validating the assumption that children have learned to read in your schools. (This is a final step in a simple-minded and inexpensive set of inquiries into the assumption. There are many more ways in which to test the assumption -- complex and highly technical and terribly expensive ways that I won't even mention.) As my final step, I would ask the children -- the learners themselves. I would make a point of asking the same sixth-graders who have just demonstrated 93% (or whatever) literacy as a group. That is, I wouldn't ask older kids or younger kids or different kids; I would ask the same kids who took the literacy test. ("Hey... you, crummy kid, where'd you learn to read?") Actually, I would plan the asking procedure with at least as much care as I used in planning the test and developing the questionnaire for parents. Having tried this a few times, with only modest success -- in fact, mostly ending in dismal failure -- if I had to do it again, I think I'd try a method something like this:

As a first step, I'd ask a few of the children themselves, bright and schoolish and sympathetic ones, how they would advise me to go about getting this information from other kids. The procedures that children recommend are almost invariably insightful and kind. Most of the youngsters who succeed in school are, by the sixth grade, very aware of the personal circumstances that affect the school learning of other kids. I think I'd use the suggestions of the kids to formulate a short set of quite general questions aimed at finding out what helped or hindered each child in learning to read -- and then I'd use this set of questions as a guide during an individual interview with as many sixth graders as I have time to reach.

I could hear a collective drawing in of breath when I said that in the presence of measurement specialists. (I didn't mention that an attractive alternative for me is that of having student interview teams -- in the hope of getting around the ever-present fact that children who have been trained for six years to say back to the school establishment what they think the establishment wants to hear have become proficient at it. A sixth grade kid being questioned by a teacher or a counselor or some researcher from downtown will try pretty hard to tell such a questioner what he thinks the questioner wants to hear. An alternative that needs to be tried is that of training as interviewers, a few age-mates who appear to have the respect and trust of the children being interviewed -- not youngsters who have the respect and trust of the teachers, necessarily, but those who have the respect and trust of the students to be interviewed. My hunch is that the students will deceive and dissemble and role-play less in an interview conducted by a respected peer than in an interview by an adult.)

I can't say a whole lot to defend the interview as a data-gathering procedure. All sorts of things are wrong with it; I know what many of them are, as do you. I also know that these same things are wrong with written questionnaires and checklists and other less direct means for obtaining information. So, in spite of its short-

comings in the technical requirements of measurement as we practice it today, I feel strongly that asking the learner, in a direct and face-to-face encounter, probably is the most valid means we have for finding out how and why a child has learned something. If the statistics for summarizing and interpreting information gathered in this way do not now exist, it is time that they were invented -- and I don't mean statistics that will convert complex narrative information into the same shibboleths that have grown up around standardized testing, code numbers or code words that pretend to reduce subtle information about a human being into a single index that can be written on the head of a pin. (That was a digression, brought on by a fit of passion.)

I have deliberately omitted some significant elements of my topic -- like partialing out and evaluating the impact of the teacher -- but I approach my summary and conclusion. Now that the citizens who pay for public education have begun to ask for an accounting, we specialists in measurement had better look beyond our finely-honed traditions and devise new ways in which to describe the impact of the educational process. Using the Dyer student-change model as the theoretical framework within which to hypothesize further with you, I have suggested some first, quick, inexpensive, and relatively easy ways of collecting information about a limited kind of reading skill at one grade level -- in three of the four categories Dyer thinks are necessary: input, output, surrounding conditions.

Lacking both the time and the ability to do so creatively, I have not even touched on the problem of data reduction in an information system as complex as this one gets to be. It is multifarious as well as multivariate, but within the capacity of known mathematical procedures and existing computers. Even if I had the time and ability to comment on data-reduction problems in educational accountability, I think I would have subordinated that topic to the problem of information-collection. Information-collection, once it gets beyond the sterile administration and scoring of multiple-choice tests, has to do with eliciting human responses within an affective environment. In truth, this kind of information collection is a political activity and possibly should be left to investigative reporters or to school anthropologists rather than to us dried-up psychometrists. We have about run out our string with "objective" tests, I think -- gone as far as we can go, even in so academic a thing as curriculum evaluation -- and must add to our old-fashioned techniques some new procedures that use different definitions of validity and reliability (i.e., truthfulness) and that open up new fields for development in humane interpretations of human behavior.

Measurement has a great deal to contribute in the ferment that finds the schools having to "account" for their impact at the same time they are trying new ways of teaching and new teaching content. Norm-based tests are traditional among us and will continue to be useful

for a while -- as the newer kinds of criterion-related measures become more sophisticated and as observations by teachers and parents and learners acquire their own statistical handles. Unless we specialists trap ourselves into thinking that the new measurement always has to conform with the format limitations of the old measurement -- paper and pencil tasks, multiple-choice questions, mechanical or objective scoring, numerical index interpretations, and all that -- I am confident that we can relate our work to the exploding curriculum in a useful way.

J.E.D.

REFERENCES

- 1) Broudy, Harry S. The Real World of the Public Schools. (Harcourt Brace Jovanovich, 1972)
- 2) Dyer, Henry S. How to Achieve Accountability in the Public Schools. (Phi Delta Kappa, Bloomington, Indiana, 1973)
- 3) Illich, Ivan. Deschooling Society. (Harper and Row, 1971)
- 4) Krumboltz, John and Helen. Changing Children's Behavior. (Prentice-Hall, 1972)
- 5) Leonard, George B. Education and Ecstasy. (Delacorte Press, 1968)
- 6) National Center for Health Statistics. Development of the Brief Test of Literacy. (Public Health Service Publication No. 1000, Series 2, No. 27. Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. 30 cents)

.....